"You wouldn't celebrate September 11" - testing a dual-pathway model of online polarisation on YouTube

Ana-Maria Bliuc[1], Laura G. E. Smith[2], Tina Moynihan[3]

[1]University of Dundee

[2]University of Bath

[3]Western Sydney University

Paper accepted for publication in Group Processes and Intergroup Relations on 24th June 2020

Abstract

Online communication is increasingly associated with growing polarisation in society. In this research, we test a dual-pathway model of online polarisation via intergroup and intragroup interaction of supporters of opposing ideological camps on YouTube. The interaction occurs over a video parody promoting a campaign to change the date of Australia Day celebrations, a divisive issue entailing contrasting narratives about Australian identity, meanings of the Australia Day, and interpretations of colonial history. To capture ideological polarisation, we conducted computerised linguistic analysis of polarised talk in the form of comments and replies (N=1,027) from supporters and opponents of the campaign. The indicators used to capture polarisation are social identification, position certainty, and psychological distance (as reflected by increased anxiety and hostility). Our results show that most polarisation (in the form of increased hostility) occurs in conditions of expression of outgroup dissent (the intergroup interaction pathway) and the most debated content on the online forum revolves around themes relevant to group identity. In addition to contributing to the understanding of group process in an online context, another key contribution of this research is providing a theory-driven method and blueprint to detect polarisation in social media data.

"You wouldn't celebrate September 11" - testing a dual-pathway model of online polarisation on YouTube

We live in an increasingly polarised world, where contrasting narratives about social reality divide societies into opposing (and often mutually exclusive) ideological camps. Divides driven by support for different political parties (political polarisation) are becoming less relevant, while issue-based polarisation which transcends political affinities seems to be a key driver of the current state of fragmentation in Western societies - as for example in Britain, it is the people's positions on Brexit rather than political party identification that drives the polarisation in the country (Duffy, Hewlett, McGrae, & Hall, 2019). This societal fragmentation is replicated in the online domain (e.g., Facebook and Twitter) where the tendency to cluster according to ideological preferences seems to be even more accentuated (Bakshy, Messing, & Adamic, 2015, Barberá, 2015; Gaines, & Mondak, 2009).

We propose that polarisation between groups from opposing ideological camps, in the specific context of online communication, can be understood as driven by both intragroup and intergroup mechanisms, whereby ingroup members interact online with other ingroup members, and with outgroup members, respectively. In this paper, we test the proposition that these different types of online interactions represent two different pathways to polarisation. The *intragroup* pathway is exemplified in recent studies that have focused on the dynamics of echo-chambers in the online domain, and which emphasise the impact of ingroup interactions between group members (Himelboim, McCreery, & Smith, 2013; Quattrociocchi, Scala, & Sunstein, 2016). However, the internet also provides opportunities for people to interact with others who are outside their own ideological echo-chambers. This latter type of interaction represents the *intergroup* pathway. In the current research, we use social interaction data extracted from YouTube to study the effects of both intragroup and intergroup online interactions on polarisation.

We refer to polarisation in terms of *ideological polarisation*, a concept that incorporates aspects of political polarisation (defined as an expanding ideological gap between groups and increased interpersonal separation between opponents, see Harel, Maoz, & Halperin, 2020) and affective polarisation (a construct derived from social identity theory, to explain partisan divides as underpinned by an increasing divergence in affect towards the ingroup and outgroup, Iyengar & Westwood, 2015; Tajfel, 1970; Tajfel & Turner, 1979). Thus, we draw upon on a view of polarisation as rooted in ideological conflict but driven by social identity processes. By doing so, we argue that polarising online debates emerge around themes that are fundamental to the self-definition of the respective groups in conflict.

Contrasting Narratives Underpinning Polarisation in Opposing Ideological Camps

At a basic level, many forms of group behaviour can be understood as attempts by group members to change the world according to a particular (collectively shared) narrative about social reality. Such narratives can divide societies because they propose conflicting and often mutually exclusive versions of social reality, as for example in the contrasting narratives driving the climate change divide (Bliuc, McGarty, Thomas, et al., 2015). Collective narratives can be understood as coherent stories about social reality that are congruent with particular sets of values and systems of beliefs, reflecting alternative world views about how the world should be and effectively dividing the society into opposing ideological camps (Bliuc, McGarty, Reynolds & Muntele, 2007; McGarty, Bliuc, Thomas, & Bongiorno, 2009). Such narratives can help us understand consensus within a group, that is, the "truth" upon which group members agree (Bessi, Coletto, Scala, Caldarelli, Quattrociocchi, 2015). They can inform our understanding of the core beliefs and values of a particular group or, put differently, they can shape the content of a particular social identity with its associated values, beliefs, and behaviour-prescribing norms (Livingstone & Haslam, 2008).

While contrasting narratives can form the bases of societal fragmentation within opposing ideological camps (at the intergroup level), they can also provide the bases of consensus and unification within a group (at the intragroup level). Here, in seeking to understand how polarisation occurs in the context of intergroup conflict between people from opposing ideological camps, we focused on the contested understanding and the associated narratives about the Australian national identity. A key issue that Australians from opposing ideological camps debate about online is the celebration of Australia Day on January 26. The January 26 commemoration marks the landing of the first British fleet to Australian shores in 1788. There is one narrative that identifies the arrival of the fleet led by Arthur Phillip as an invasion legitimatised by the British imperialist rule, representing the beginning of the dispossession and genocide of the Indigenous people in Australia (Maddison, 2012). Supporters of this narrative advocate for changing the date of the national day via a campaign known as "Change the date".

An alternative narrative (endorsed by the current Australian government) construes this particular date as the day when modern Australia was founded (through colonization nevertheless) and advocates for maintaining the date as it is. These two narratives are also aligned to contrasting understandings and definitions of the Australian national identity (Bliuc, McGarty, Hartley, & Muntele, 2012). In particular, people who identify as Australian may support different versions of this social category - on one hand, a nationalistic version with its core values linked to race, ethnicity, and religion (i.e., British, white, Christian) aligned to support for maintaining the current date to celebrate Australia's day, and choosing to overlook the significance of the date for Indigenous Australians. On the other hand, Australians may identify with a more inclusive version of the national identity (Reicher, Cassidy, Wolpert, Hopkins, & Levine, 2006), supporting cultural diversity and more sensitive towards issues pertinent to Indigenous Australians. This version of national identity is aligned

to support for a change of date (Maddison, 2012; Moran, 2011; Pakulski & Tranter, 2000). Thus, support or opposition to the "Change the date" campaign cannot be solely explained by reference to social categories based on political orientation or cultural belonging. Instead, in this context, we can conceptualise the intergroup conflict driving the debate as conflict between two opposite ideological camps based on shared views about what being a "true" Australian means (Bliuc et al., 2012; Reicher et al., 2006).

Online Ideological Polarisation on YouTube

Rapid advances in information and communication technologies have enabled us to live highly connected lives in which finding like-minded people – those who share our views about the world – is only a few clicks away. As a result, people can choose to primarily interact online with others who share their views, or when engaging with social and political issues, with those who belong to the same ideological camp as them (social interactions at an intragroup level). Where people are only exposed to information from the ingroup, *echo-chambers* can form (Bakshy et al., 2015). Platforms such as Twitter and Facebook, where exposure to similar ideas is embedded in their core functionality, provide ideal conditions for the formation of echo-chambers (Barberá, Jost, Nagler, Tucker, & Bonneau, 2015; Himelboim et al., 2013) and further polarisation (Barberá et al., 2015; Bessi et al., 2015; Conover et al., 2011).

According to classic research in social psychology, polarisation can occur via increased *intragroup interaction* where group discussion on contentious issues results in division within the group and a shift to the extreme group positions (Moscovici & Zavalloni, 1969). In other words, group polarisation occurs when members of a deliberating group move towards a more extreme point in direction of the initial position of the group, so that, for example, members of anti-abortion group discussing the issue of abortion, will become more extreme in their views after the discussion (Sunstein, 1999; see also Myers & Bishop, 1970).

The same polarising effect can be observed in online interactions occurring in echo-chambers. In echo-chambers, beyond the shift in the initial position, polarisation seems to lead to a decreased likelihood to interact with outgroup members (Quattrociocchi, Scala, & Sunstein, 2016) and the formation of "identity bubbles" (Kaakinen, Sirola, Savolainen, & Oksanen, A., 2020; Oksanen, Oksa, Savela, Kaakinen, & Ellonen, 2020). The polarisation of attitudes in echo-chambers is well-illustrated by research on the far-right in which online participation in radical echo-chambers (such as neo-Nazi and white supremacist forums) lead to changes in group norms and increased unification at intragroup level (Bliuc, Betts, Vergani, et al., 2019; 2010; Bliuc, Betts, Faulkner, et al., 2020; Wojcieszak, 2010; for a review see Bliuc, Faulkner, Jakubowicz, & McGarty, 2018).

While *user-driven* spaces with restricted/members-only forums such as Facebook and discussion groups are best suited to encourage echo-chambers and therefore intragroup interaction, *content-driven* sites such as online open discussion sites are ideal platforms for intergroup communication. In particular, there are many (content-driven) online spaces that encourage interaction between people from opposing ideological camps. Such online spaces post contentious content which is debated by online users who identify with a particular ideological camp (or opinion) in relation to the issue. For example, an online news article on the pro-life movement would likely draw users who identify either as pro-life or pro-choice, so that its comments section would represent a platform for both intragroup and intergroup interaction (i.e., interactions between both users who agree and between users who disagree on the issue of abortion).

In echo-chambers, polarisation occurs because of both intragroup influence and restricted access to diverse views (political homophily, Vaccari, Valeriani, Barberá, et al., 2016). However, exposure to diverse views does not necessary represent an antidote to polarisation. On the contrary, exposure to diverse and in particular opposing views of an

ideological outgroup can be seen as another pathway to polarisation – that is, polarisation via *intergroup interaction* (Bail, Argyle, Brown, et al., 2018; Pauwels & Schils, 2016). When interacting with people with opposing views, polarisation occurs through a process of "disconfirmation bias" (see Karlsen et al., 2017; Taber, Cann, & Kucsova, 2009) when people become stronger in their beliefs after being presented with counterarguments from an outgroup.

Indicators of ideological polarisation

Our understanding of polarisation is drawn from definitions of political and affective polarisation and can be operationalised as both *ideological and psychological distancing* between people from opposing ideological groups (see also similar conceptualisations of opinion polarisation from physical statistics, Castellano, Fortunato, & Loreto, 2009). Social interaction in conditions of ideological intergroup conflict can result in polarised social identities (including in online contexts, see Pauwels & Schils, 2016) because the groups are each fighting for their own cause – emphasising and exacerbating intergroup differences (Bliuc et al., 2012; 2015). This is because, according to the social identity approach (SIA, Tajfel & Turner, 1979; Turner et al., 1987), in contexts of intergroup conflict group identities become more salient, a process which in turn leads to group members' adoption of stereotypical group behaviour and expression of social identification (Turner et al., 1987). This *ideological distancing* will therefore manifest as increased identification with the respective group (Bliuc et al., 2015) and increased certainty and confidence about the group position (Bliuc et al., 2007). In ideological camps defined by a particular position which is in opposition to the outgroup, disagreement with the outgroup is expected, so intergroup interaction should further increase group identification via increased position certainty (Feixas, & Winter, 2019; Holtz & Miller, 2001; Holtz & Nihiser, 2008; Turner & Tajfel, 1986). Psychological distancing, on the other hand, as increasing opposition between ingroup

and outgroup, can be captured by collective emotions such as hostility and intergroup anxiety, which can arise as a result of perceived threat (in this case, symbolic) from interaction with outgroup members (Stephan & Stephan, 1985, Stephan & Renfro, 2002; Stephan, Ybarra, & Morrison, 2017). In this context, the construct of psychological distance applies to intergroup processes, therefore differing from alternative conceptualisations where psychological distance refers to an individual process of affective detachment where people tend to 'psychologically remove themselves from emotionally painful events' (Cohn, Mehl, & Pennebaker, 2001; Pennebaker & King, 1999).

Present Research

In the current research, we aim to address the question of what happens in online spaces when people interact outside their echo-chambers, that is, in virtual spaces where they can interact with both people who share their views and people who have opposing views to their own. We seek to test the effects of both intragroup and intergroup interaction on polarisation, therefore, we used YouTube to examine social interactions occurring at both levels. YouTube is particularly well suited for this type of inquiry as it is a platform that by design can engage users from opposite ideological camps through the distribution of controversial socio-political video content. Unlike social networking sites such as Facebook and Twitter where the focus is on the relationships between users, YouTube is focused on content viewing (de Bérail, Guillon, & Bungener, 2019; Khan, 2017), with the option to react to the content and also interact with other users by commenting and replying to comments of the other users.

The main objective of our study is to test a dual-pathway model to ideological polarisation in an online context; a first pathway to polarisation being via intragroup interactions with ingroup members (i.e., similarly to processes occurring within echo-

chambers), while a second pathway being via intergroup interactions with outgroup members. We investigate these pathways in an online context where online users from opposing ideological camps are exposed to both ingroup consensus and outgroup dissent (Grauwin & Jensen, 2012; Liu & Srivastava, 2015; Burke & Goodman, 2012; Guo & Harlow, 2014; Harlow, 2015). In our study, we captured polarisation by focusing on three key dimensions of polarised talk: a) social identification with one ideological camp; b) certainty (of the ideological position of the respective camp); and c) psychological distance from the opposed ideological camp (increased anxiety, hostility). More specifically, we sought to test the general prediction that those dimensions of polarisation are increased by both online social interactions between people from opposing camps (i.e., interaction in the context of *intergroup dissent*) and online interaction with people from the same camp (i.e., interaction in the context of *intragroup consensus*). Furthermore, intergroup conflict between groups drawn from opposing ideological camps is likely to be an identity-driven process where issues with direct relevance to the group's self-definition are the most debated. From this, it follows that the most polarising topics that the opposing sides of a conflict engage with would likely be most directly pertinent to group's identity.

However, different indicators of group polarisation would be differently affected by intragroup versus intergroup interactions. In particular, it was expected that intragroup interaction (interaction in the form of intragroup position validation and endorsement) would primarily lead to increases in position certainty or self-defining belief validation (H1), while intergroup interactions (interaction in the form of expressing and endorsing intergroup dissent) would most likely lead to increases in social identification (as captured by the use of pronouns) and psychological distance as captured by anxiety, hostility and use of swear words (H2). In addition, it is expected that the most debated themes on the online forum will

be pertinent to group identity, that is, polarisation will mostly occur in clusters debating these themes (H3).

By focusing on online contexts of interaction, we can test this *dual-pathway model of polarisation* where polarisation occurs in the conditions of intergroup conflict via online social interaction *within* and *between* opposite ideological camps. In other words, we examine interaction at *intragroup* level (in the context of consensus within the ingroup) and *intergroup* level (in the context of dissent with the outgroup) by focusing on social interactions between two groups drawn from opposing ideological camps on YouTube. The specific issue that is debated here is about changing the date for the Australia Day (i.e., supporting versus being opposed to the "Change the date" campaign). However, the real basis of division that drives the polarisation in this case is conflict about competing narratives about Australian national identity. This conflict divides the Australian society into opposing ideological camps which are distinct psychological groups that endorse alternative narratives about the meaning of the Australian national identity with corresponding contrasting interpretations of historical events.

Method

Data collection

We extracted anonymous comments and replies to comments that were posted in response to an Australia Day parody video "Change the date of Australia Day" (http://www.youtube.com/watch?v=UytdM-x3cv4) on YouTube. The video was produced by The Juice Media (http://www.thejuicemedia.com.au), an independent online media website and posted on YouTube on 24/01/2017. The video is a parody of anti-piracy video popular in Australia in the 2000s (http://www.youtube.com/watch?v=HmZm8), and makes parallels between Australia day commemorations and celebrating historical disasters and atrocities such as September '11, Hiroshima and the Holocaust (i.e., historical dates of human atrocity).

(The Juice Media, 2017)

As of March 2020, the video had a total of 155,374 views and generated 4.3K likes and 4.1K dislikes, 1 million views and 9K shares on Facebook. Data was collected and analysed at the level of contribution, not the contributor. From the total of 1,415 contributions (comments and replies) from 646 users, only the comments and replies that could be categorised as either in favour or against the campaign were retained for the analysis (a total of 1,027; *N* comments = 434 and *N* replies = 593). Seventy-six comments were categorised as being in favour of the "Change the date campaign" (contributed by the supporters of the campaign) and 358 as being against the campaign (contributed the opponents of the campaign), while 265 replies were in favour of the campaign and 328 against.

Coding process. Two of the researchers conducted the coding of the comments and posts as either being in favour or against the 'Change the date' campaign, an approach drawn from previous research on racism that used similar categorisations of online posts (Bliuc et al., 2012; Faulkner & Bliuc, 2016). First, the criteria of categorisation were discussed to ensure that both coders had the same understanding of the categorisation criteria. That is, the posts (comments and replies) were assessed as to whether they contained either

agreement/expressions of positive attitude or disagreement/expressions of negative attitude towards:

a) the campaign (e.g., pro-campaign content: “totally agree with this. This is why i don't celebrate Australia day (…)” versus anti-campaign content: “Absolute crap. Throw some lamb on the barbie and shut the f** up.”; “Preachy and ineffectual. Just like the original ad”);

b) the particular historical narrative endorsed by each of the two camps (e.g., pro-campaign content: “What you're celebrating is mass murder of traditional owners of this land. You owe it to yourselves to better educate yourself of Australia's past (…)” versus anti-campaign content: ““f** you this was the day cap cook landed not the day we colonised”);

c) the members of the outgroup (e.g., pro-campaign content: “You have angered the mindless nationalists. Keep up the good work. They only respond like this to that which they fear - namely the education of their peers.” versus anti-campaign content: “You hate Australia Day so much? That terrible day when those nasty Brits arrived? Fine, get rid of those clothes and put a loincloth back on. Centrelink payments? Stop taking those too. Get rid of your cars, stop using all that lovely medicine, abandon your houses and go back deep into the bush where you can live as your uncivilised people once did, you wouldn't last two minutes.”)

Initially, about 10% of all posts were coded in a joint session that allowed the coders to discuss posts they were uncertain about and agree on the categorisation. This session was followed by independent coding by each of the coders of the remaining posts. The process concluded with a final session where the codes were reviewed and any discrepancy in coding resolved. At the end of this process, some of the posts were excluded from the analysis, that is, posts including either content about the artistic quality of the video (or the lack of it) or ambiguous content unrelated to the campaign.

Study design

The naturalistic context of the online interaction in the form of comments and replies to the YouTube video advocating for the 'Change the date' campaign provided ideal conditions (harder to achieve in a traditional lab environment) to test how indicators of polarisation are affected by both intragroup and intergroup interaction. In particular, comments made in response to the video could be classified as either in favour of the campaign (i.e., context of intragroup validation/expressing consensus) or against the campaign (i.e., context of intergroup contestation/expressing dissent). Replies made in response to comments could be also classified as either in favour or against the campaign, but the possibility to express agreement to comments against the campaign provided an additional level of interaction – going beyond expressing consensus or dissent towards the campaign, there is the possibility of both *endorsing dissent* (replies opposing the campaign) and *endorsing consensus* (replies in favour of the campaign). Therefore, the structure of the online interactions provided conditions for four types of interactions on the YouTube video:

1. Expressing consensus (direct intragroup validation: comments in favour of the campaign)
2. Expressing dissent (direct intergroup contestation: comments against the campaign)
3. Endorsing consensus (indirect intragroup validation: replies in favour of the campaign)
4. Endorsing dissent (indirect intergroup contestation: replies against the campaign)

This created four types of interactions, and two independent variables for our analyses: group categorisation (by position towards the campaign: pro-campaign/position validation versus anti-campaign/position contestation) and type of contribution (comment versus reply).

The dependent variables were the indicators of polarisation in the form of identification, position certainty, and psychological distance (captured by anxiety, hostility

and use of swear words). To identify the most polarising themes in the online forum, we identified the most debated comments (those comments that generated the most replies).

Measures

We used LIWC software (Pennebaker, Booth, Boyd, & Francis, 2015) to process the textual data in the posts and create quantitative variables to capture polarisation. The software automatically scans the chosen text (i.e., the comment or reply) for words and phrases that are present in the LIWC 2015 dictionary (lexicon) and provides a score to indicate the percentage of words in that text that are in specific dictionary word categories. The word categories include both linguistic categories (e.g., pronouns) and psychological dimensions (e.g., affect, cognition, drives, etc.). LIWC has been used in numerous studies to measure various variables, with different samples and has shown that the psychometrics of word categories are valid (for a review see Tausczik & Pennebaker, 2010).

Social identification. To capture social identification, we used the LIWC subcategories of first-person plural pronouns (e.g., we, our, us), and third person plural words (e.g., they, their, they'd) to capture social identification with the respective camps. The use of words that capture intergroup ("us versus them") language as indicating social identification is consistent with previous research (Smith, Gavin, & Sharp, 2015; e.g., Bliuc et al., 2017; 2018; 2019; Faulkner & Bliuc, 2018).

Position certainty. The LIWC categories of certainty as a positive indicator (e.g., always, never) and tentativeness as a negative indicator (e.g., maybe, perhaps) were used to capture certainty.

Psychological distance. Three indicators of psychological distance were used: anxiety, anger, and use of swear words (the latter two to capture hostility). We used the LIWC category *anxiety* (e.g., worried, fearful). For hostility, we used the LIWC categories *anger* (e.g., hate, kill, annoyed), and *swear words* (e.g., f***, damn, sh**).

Data analysis

The structure of online interactions between users on YouTube creates conditions for four types of interactions, and two independent variables for our analyses: group categorisation (by position on the campaign: pro-campaign/position validation versus anti-campaign/position contestation) and type of contribution (comment versus reply). Means, standard deviations (SD), and correlations were calculated for all variables. Mean differences in the indicators of polarisation across the four types of interactions were compared using the open-source statistical software JASP (JASP Team, 2020).

In addition to the quantitative analysis, to determine whether the most debated themes on the online forum are pertinent to group identity (H3), the text of comments that generated the most replies were collated and analysed using an approach derived from thematic analysis (based on the steps outlined by Braun & Clarke, 2006). In particular, the analysis of the qualitative data was structured into several steps as illustrated in Table 1:

Insert Table 1 about here

Results

Preliminary analyses

Across both comments and replies, the mean number of words for contributions made by supporters and opponents of the campaign were similar (*N* supporters = 341 , *M* = 54.196, *SD* = 82.478, range from 1 to 806; and *N* opponents = 686, *M* = 50.063, *SD* = 69.135, range from 1 to 782 for opponents). The descriptive statistics (means, standard deviations and Pearson's correlations) for all the dependent variables (i.e., we, they, certainty, tentativeness, anxiety, anger and swear words) are shown in Table 2.

Insert Table 2 about here

Main analyses

The effects of type of social interaction on polarisation. We conducted MANOVA to test whether intragroup and intergroup interactions resulted in relative differences in polarised talk, indexed by expressions of ingroup identification (as captured by the use of 'we' and 'they' pronouns), certainty (as captured by the use of certainty and tentativeness linguistic categories), and psychological distance (as captured by the use of anxiety, anger, and swear words categories).

In the MANOVA, we entered group categorisation (by position on the campaign: pro-campaign/position validation versus anti-campaign/position contestation) and type of contribution (comment versus reply) as independent variables. The assumption of homogeneity of covariance was violated (Box's $\chi^2 = 1549.700$, $p < .001$), so we report Pillai's trace values in addition to Wilks Λ. The analysis shows a significant multivariate effect for the interaction between group categorisation and type of contribution (Wilks $\Lambda = .969$, p <.001; $\text{Trace}_{\text{Pillai}} = .031$, $p < .001$). Both group categorisation and the type of contribution had significant separated effects on the dependent variables (Wilks $\Lambda = 0.977$, $p = .001$, $\text{Trace}_{\text{Pillai}} = .023$, $p < .001$ and Wilks $\Lambda = 0.943$, $p < .001$, $\text{Trace}_{\text{Pillai}} = .57$, $p < .001$, respectively). As a follow-up to MANOVA, a series of one-way ANOVAs were conducted on each of the dependent variables. Means and standard deviations both across comments and replies and across pro-campaign and anti-campaign contributions are reported in Table 3. F-values with corresponding effect sizes for all the indicators are reported in Table 4.

Insert Tables 3 and 4 about here

Post-hoc analyses were conducted to examine individual mean difference comparisons across the groups and different types of contributions. The assumption of homogeneity of variance was violated for most of the variables, so unless specified otherwise, Games-Howell post-hoc comparisons are used. We found a mixed pattern of

differences in levels of all the indicators of polarisation across groups and different types of contributions with the exception of anxiety (which did not significantly vary across groups and types of contributions).

Identification. No significant differences were found between levels of use of first person plural pronouns (e.g., "we") across contributions from the pro-campaign versus anti-campaign groups. However, the use of these pronouns was significantly higher in comments than in replies (M_{diff} = 0.427; t = 2.227, p_{tukey}=0.026, Cohen's d = 0.142), suggesting that online contributors tended to focus on the ingroup (rather than the outgroup) when expressing their ingroup opinion. Significant differences were found in the use of third person plural pronouns (e.g., "they"), which was significantly higher in replies compared to comments (M_{diff} = -0.477; t = -2.742, p_{tukey}=0.006, Cohen's d = 0.176), suggesting that references to outgroup (as "they") were higher when the ingroup position was endorsed (in replies), rather than when group position was expressed (in comments). However, there was no significant simple interaction effect found for both the use of "we" and "they" pronouns.

Position certainty. Certainty tended to be higher in posts from the pro-campaign group compared to the anti-campaign group, but the difference did not reach significance (M_{diff} = 0.428; t = 1.856, p_{tukey}<0.064, Cohen's d = 0.128). There was no significant difference between levels of certainty in comments versus replies. In the case of tentativeness (as a negative indicator of position certainty) the assumption of homogeneity of variance was met, so we report results of standard post-hoc comparisons. Tentativeness was higher in replies compared to the comments (M_{diff} = -0.678; t = -2.451, p_{tukey}=0.014, Cohen's d = 0.097), but they were no significant differences between levels of tentativeness across groups (M_{diff} = -0.530; t = -1.917, p_{tukey}=0.056, Cohen's d = 0.046). One simple interaction effect was significant, showing that tentativeness in the pro-campaign comments was significantly lower than in anti-campaign replies (M_{diff} = -1.208; t = -2.589, p_{tukey}=0.048, Cohen's d =

0.346). These findings suggests that in direct expressions of group position (in comments), there are lower levels of tentativeness that in replies, when the group position is endorsed, and this effect is maintained in expressions of group position in the pro-campaign camp (in comparison to when anti-campaign group position is endorsed in replies).

Anxiety. For anxiety, the assumption of homogeneity of variance was met, so we report results of standard post-hoc comparisons that show that anxiety did not significantly differ between comments and replies (M_{diff} = 0.079; t = 0.874, p_{tukey}= 0.382, Cohen's d = 0.034), and between pro-campaign and anti-campaign contributions (M_{diff} = 0.022; t = 0.239, p_{tukey}= 0.811, Cohen's d = 0.021). No interaction effects were significant.

Hostility. Anger was higher in contributions from the anti-campaign group (M_{diff} = -1.716; t = -4.222, p_{tukey}<0.00, Cohen's d = 0.254) and higher in comments than in replies (M_{diff} = 3.002; t = 5.976, p_{tukey}<.001, Cohen's d = 0.395). Three simple interactions were also significant. That is, anger was significantly higher in anti-campaign comments compared to pro-campaign comments (M_{diff} = -.2.861; t = -3.146, p_{tukey} = 0.009, Cohen's d = 0.349), in anti-campaign comments compared to pro-campaign replies (M_{diff} = 3.530; t = 6.050, p_{tukey}<.001, Cohen's d = 0.499), and in anti-campaign comments compared to anti-campaign replies (M_{diff} = 3.481; t = 6.326, p_{tukey}<.001, Cohen's d = 0.441). The use of swear words was higher in comments than in replies (M_{diff} = 3.354; t = 5.986, p_{tukey}<.001, Cohen's d = 0.399), and in contributions from the anti-campaign group compared to those from the pro-campaign group (M_{diff} = -1.838; t = -4.300, p_{tukey}<.001, Cohen's d = 0.253). Again, three simple interactions were significant, the use of swear words being significantly higher in anti-campaign comments compared to pro-campaign comments (M_{diff} = -32.565; t = -2.676, p_{tukey} = 0.038, Cohen's d = 0.418), to pro-campaign replies (M_{diff} = 3.908; t = 6.139, p_{tukey}<.001, Cohen's d = 0.282), and anti-campaign replies (M_{diff} = 3.746; t = 6.239, p_{tukey}<.001, Cohen's d = 0.442). These findings show that intergroup interaction (and in particular in the conflict

conditions present when anti-campaign expressions of group position were made) resulted in increased psychological distance as captured by increased hostility in polarised talk.

Thematic analysis of the most debated comments

To examine H3, we analysed the content of comments that received 5 or more replies, that is, from a total of 434 comments only 20 received 5 or more replies. The number of replies for these comments ranged from 5 to 123 (M = 27.65, SD = 32.43). From these comments, 3 comments were in favour of the campaign and 27 against. The description of the final themes that emerged from the analysis of the most contentious comments is presented below. Two inter-related key themes were identified: one about the *interpretation of the historical narrative* regarding the arrival of the first fleet to Australia, and the second about the *contested meanings of the Australia day*. In line with our predictions, both these themes were debated by using identity-related arguments that were promoting contrastive versions of the Australian identity in a similar way found in previous research on the differing meaning of social identity content (Bliuc et al., 2012; Reicher et al., 2006). Table 5 summarises the key themes emerging from the analysis of the most debated comments together with the specific arguments used for each and corresponding illustrative quotes.

Insert Table 5 about here

Most narratives about the historic truth were based on the argument that no harm was done to the Indigenous population on that specific day, but this version is contested in contributions that suggest that it does not really matter what happened on that day – the reality is that it marked the beginning of a period of injustice and violence against the Indigenous population. More extreme versions of this narrative see Australia as a country that was conquered through a war which was lost by the original inhabitants (see Extract 2 in Table 5).

The meaning of Australia day on that particular date was construed in contrasting narratives as either marking the start of a period of grave injustice and violence against Indigenous Australians or marking the start of modern Australia, a wealthy and successful nations (thus being deserving of celebration). This narrative in particular seemed to be linked to the most blatant racist content against Indigenous Australians (in arguments that suggested that as recipients of welfare, Indigenous Australians have the most to gain from the modern state). A less blatantly racist version of this narrative was around calls for unity in celebrating Australia, so that the meaning of Australia Day seen as being about celebrating modern Australia not the past (see Extract 4 in Table 5).

Importantly in the context of our predictions (and in particular H3), both of these themes are directly relevant to definitions of the Australian identity and entail distinct values, norms, and emotions. That is, one of the narratives about the historical events around January 26 promotes a version of history about the roots of Australian identity where colonisers had nothing to be feel guilty about (either no violence was committed or the violence was justified). At the same time, the opposing narrative assigns blames to the colonisers, but at the same time extends the definition of what it means to be Australian by including the Indigenous population. A similar process could be seen in the case of the contrasting narratives about the meaning of the celebration. Again, one narrative promoted an exclusive focus on the present and future of the country, so Australia day was construed as a celebration of the benefits of modern Australia, while ignoring the past (a positive image of Australian identity is endorsed). The opposing narrative was more focused on the past and promoting reparation for the injustice and harm done to Indigenous Australians by the British colonisers (this narrative recognised the past wrongdoings and proposes a version of national identity that acknowledges guilt and apology).

Discussion

The key aim of this study was to test a dual-pathway model of polarisation that would capture how ideological polarisation can occur both via intragroup interaction (position validation by ingroup) and intergroup interaction (position contestation by outgroup). Specifically, we expected that intragroup interactions would mostly increase confidence in the group position (position certainty, H1), while intergroup interactions would mostly increase ingroup identification (via increased social identity salience due to the intergroup context and increased group-defining position certainty) and group differences resulting in increased psychological distance (as captured by anxiety, hostility, and use of swear words, H2). Interaction was conceptualised as online communication between contributors from opposing ideological camps, so it was also expected that, in terms of the content of communication, the most debated themes would be relevant to the group's identity (H3). Thus, by analysing the most debated content in the online forum, we expected to be able to detect specific identity-relevant markers of polarised talk in this specific context.

These propositions were tested in opposing ideological camps formed around contrasting narratives about national identity in Australia. That is, we examined differences in polarised talk (captured by linguistic indicators of identification, certainty, tentativeness, anxiety, and hostility - anger and swearing) across different types of social interactions about the 'Change the Date' campaign in Australia. Our argument was that public division driven by this campaign represents more than a conflict of opinions about when to celebrate the Australia day; rather, in line with previous work, it encapsulates conflict between psychologically meaningful social identities that are aligned to particular social categories, but are not reducible to these (Bliuc et al., 2015; 2019).

The online context of social interactions in response to a YouTube video supporting the Change the date campaign in Australia provided ideal conditions to test our theoretical assumptions as captured by the effects of naturalistic interactions between supporters and

opponents of the campaign on the indicators of ideological polarisation. That is, comments from supporters of the campaign made directly in response to the video represented intragroup validation (expressing group position as intragroup consensus) while comments from opponents of the campaign represented intergroup invalidation (expressing group position as intergroup dissent). The way in which YouTube is designed to provide opportunities to users to engage with comments, created an opportunity for us to examine the effects of conditions in which there is an indirect expression of intragroup validation and intergroup invalidation. In particular, replies in favour of the campaign were used to endorse consensus (indirect intragroup validation), while replies against the campaign to endorse dissent (indirect intergroup invalidation). Thus, the nature of the structure of this online social interaction data enabled us to compare different pathways to polarisation and their differential effects on polarised talk in real supporters and opponents of a campaign addressing core defining issues about what a modern Australian national identity may entail. Our study proves that online platforms can play an important role in theory testing that goes beyond the lab and applies to issues that really matters to people and society at large.

Our hypotheses, however, were not entirely supported by our data. That is, we found a main effect of the interaction between group and type of contributions on polarisation indicators, suggesting that polarisation is indeed affected by the type of the interactions between members of opposing ideological camps. However, the simple effects of these interactions on various indicators were mixed. In particular, we found that the indicators of hostility (anger and use of swear words) were highest in conditions of direct expression of intergroup dissent. That is, when people interacted with content that contradicted their ingroup ideology, expressions of hostility were at their highest (not surprisingly, higher than in comments and replies from the pro-campaign group, but also higher than anti-campaign replies/expression of endorsement of dissent). In line with our predictions, ingroup position

validation (in the form of decreased tentativeness only) was increased in conditions of intragroup validation (when expressing consensus with the ingroup in the form of comments in favour of the campaign). However, contrary to our prediction, group identification as captured by the use of first person and third person plural pronouns did not vary across the contributions from different groups (between supporters and opponents of the campaign) in both comments and replies, but rather a focus on 'us' as a collective entity (as reflected in the use of first-person plural pronouns) seemed apparent in conditions of *expressing* the group's ideology (versus *endorsing* it). Also, anxiety as capturing one aspect of psychological distancing, did not significantly change across conditions. Overall, we found that it is mostly intergroup interaction, when direct dissent is expressed, that drove polarisation, in particular in the form of increased hostility. Whilst this study used a relatively small sample and focused on a specific context, these findings are consistent with research from communication science on online political newsgroups reporting showing that discussion tends to happen between clusters of like-minded groups (via intergroup processes) rather within these clusters (via intragroup processes see Kelly, et al., 2005). This preference for such discussion may be driven by a motivation to deliberate, debate, and argue rather than agree with others, which is particularly salient in the case of emotionally charged and controversial socio-political issues (Yardi, & Boyd, 2010).

Our analysis of qualitative data is consistent with previous research on social identity content, showing that a single social category (e.g., Australian national identity) can incorporate different sets of beliefs, values and associated norms, and therefore division and conflict can occur as a result of contestation of the very bases of the social category (Bliuc et al., 2012; Reicher et al., 2006). Put simply, division and intergroup conflict can occur because of dissent around the meaning of a social category. Furthermore, our findings suggest that these debates around the meaning of Australian national identity in this case,

were likely conducive of further polarisation between competing groups. Ideally, this theoretical point would be further tested in studies using experimental designs (so causality can be clearly established by investigating whether debates on identity-relevant issues lead to further polarisation between members of opposing groups).

Despite the affordances provided by the online context of our study, there are several limitations imposed by this same context. For instance, our findings are based on the study of polarisation processes in the specific context of online debates about a particular political issue in a particular cultural context, that is, the legitimacy of celebrating Australia's national day on January 26. We believe, however that given that the indicators of polarised talk that we used in this study, and the theoretical basis for our hypotheses, are independent of the socio-political context under investigation, it is possible that the same processes and patterns of results can be observed in other contexts. Indeed, we argue that it is likely these findings would apply to other debates about divisive issues in society (including debates about the meaning of national identity but extending beyond these), but future studies should firmly establish if this is the case. Furthermore, in the current research, these debates are generated by exposure to one video that promotes the 'Change the date' campaign. If this research is seen as analogous with studies conducted in the laboratory, this would be the equivalent of drawing conclusions based on a single study. Therefore, we recommend that our propositions should be further tested in the future by a) using larger samples of online textual data, b) conducting further studies using similar designs, and c) using additional indicators of polarization. These recommendations would also help address the issue of relatively small magnitude of the effects that we found.

Conclusion

In sum, we found that in an online context of debates around national identity, intergroup interactions increase hostility, while intragroup interactions decrease tentativeness in expressions of group position (in comments). Our qualitative analysis of the most debated contributions suggests that identity debates are at the core of polarised talk in the context of our study. The key implication here is that there appears to be a dual pathway to polarisation but that intergroup and intragroup interactions have varied effects on different indicators of polarisation. Overall, the present study demonstrates that supporters and opponents are not just individuals who hold competing views about what it means to be Australian; rather they are individuals who belong to distinct and opposing psychological groups that endorse contrasting narratives about Australian identity. Thus, when they interact online they reproduce patterns that may occur within and between other groups in conflict (such as racist and activist groups, see Faulkner & Bliuc, 2018; Bliuc et al., 2018; 2019; Smith, Wakeford, Cribbin, Barnett, & Hou, 2020).

One of the key contributions of this paper is that we provide a theory-driven method and blueprint to capture polarisation in social media data. Considering potential applications of our findings, the current study contributes to understanding how online content about a divisive issue (in the context of Australia in this case) brings together members of opposing ideological camps who engage in different types of social interactions within their ingroup and with the outgroup. Online platforms such as YouTube support intergroup communication and provide a platform for potential (cross-group) social learning (Guilbeault, Becker, & Centola, 2018), so they can be seen as valuable opportunities for social scientists to study such communication *in vivo*.

References

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., ... & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, *115*, 9216-9221. doi:10.1073/pnas.1804840115

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*, 1130-1132. doi:10.1126/science.aaa1160

Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, *23*, 76–91. doi:10.1093/pan/mpu011

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, *26*, 1531-1542. doi:10.1177/0956797615594620

Bessi, A., Coletto, M., Davidescu, G. A., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2015). Science vs conspiracy: Collective narratives in the age of misinformation. *PloS one*, *10*, e0118093. doi:10.1371/journal.pone.0118093

Bliuc, A. M., Betts, J. M., Faulkner, N., Vergani, M., Chow, R. J., Iqbal, M., & Best, D. (2020). The effects of local socio-political events on group cohesion in online far-right communities. *PloS one*, *15*, e0230302. doi:10.1371/journal.pone.0230302

Bliuc, A. M., Betts, J., Vergani, M., Iqbal, M., & Dunn, K. (2020). The growing power of online communities of the extreme-right: deriving strength, meaning, and direction from significant socio-political events 'in real life'. *ICCT Journal*, https://icct.nl/publication/the-growing-power-of-online-communities-of-the-extreme-right-deriving-strength-meaning-and-direction-from-significant-socio-political-events-in-real-life/. doi: 10.97812345/2020.4.3

Bliuc, A-M., McGarty, C., Hartley, L., & Muntele Hendres, D. (2012). Manipulating national identity: The strategic use of rhetoric by supporters and opponents of the Cronulla riots in Australia. *Ethnic and Racial Studies, 35,* 2174-2194. doi:10.1080/01419870.2011.600768

Bliuc, A-M., McGarty, C., Reynolds, K., Muntele, D. (2007). Opinion-based group membership as a predictor of commitment to political action. *European Journal of Social Psychology, 37*, 19-32. doi:10.1002/ejsp.334

Bliuc, A-M., McGarty, C., Thomas, E. F., Lala, G., Berndsen, M., & Misajon, R. (2015). Public division about climate change rooted in conflicting socio-political identities. *Nature Climate Change, 5*, 226 – 230. doi:10.1038/nclimate2507

Bliuc, A. M., Betts, J., Vergani, M., & Dunn, K. (2019). Collective Identity Changes in Far-right Online Communities: The Role of Offline Intergroup Conflict. *new media & society, 21,* 1770-1786. doi:10.1177/1461444819831779

Bliuc, A. M., Faulkner, N., Jakubowicz, A., & McGarty, C. (2018). Online networks of racial hate: A systematic review of 10 years of research on cyber-racism. *Computers in Human Behavior*, *87*, 75-86. doi:10.1016/j.chb.2018.05.026

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences, 114*, 7313-7318. doi:10.1073/pnas.1618923114

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in Psychology, 3*, 77-101. doi:10.1191/1478088706qp063oa

Burke, S., &. Goodman, S. (2012). Bring back Hitler's gas chambers: Asylum seeking, Nazis and Facebook - A discursive analysis. *Discourse & Society, 23*, 19-33. doi:10.1177/0957926511431036.

Castellano, C., Fortunato, S., & Loreto, V. (2009). Statistical physics of social dynamics. *Reviews of Modern Physics, 81,* 591-649. doi:10.1103/revmodphys.81.591

Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding September 11, 2001. *Psychological Science, 15*(10), 687-693. doi: 10.1111/j.0956-7976.2004.00741.x

Conover, M., Ratkiewicz J, Francisco M (2011) Political polarization on twitter. *ICWSM, 133,* 89–96. doi:10.1109/passat/socialcom.2011.34

de Bérail, P., Guillon, M., & Bungener, C. (2019). The relations between YouTube addiction, social anxiety and parasocial relationships with YouTubers: A moderated-mediation model based on a cognitive-behavioral framework. *Computers in Human Behavior, 99,* 190-204. doi:10.1016/j.chb.2019.05.007

Del Vicario, M., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2017). Modeling confirmation bias and polarization. *Scientific Reports, 7,* 40391. doi: 0.1038/srep40391

Duffy, B., Hewlett, K., McCrae, J., & Hall, J. (2019). Divided Britain? Polarisation and fragmentation trends in the UK. King's College London, The Policy Institute, https://www.kcl.ac.uk/policy-institute/assets/divided-britain.pdf

Faulkner, N., & Bliuc, A. M. (2018). Breaking down the language of online racism: A comparison of the psychological dimensions of communication in racist, anti-racist, and non-activist group. *Analyses of Social Issues and Public Policy. 18,* 307-322. doi:10.1111/asap.12159

Faulkner, N., & Bliuc, A. M. (2016). 'It's okay to be racist': moral disengagement in online discussions of racist incidents in Australia. *Ethnic and Racial Studies, 39*(14), 2545-2563. doi:10.1080/01419870.2016.1171370

Feixas, G., & Winter, D. A. (2019). Towards a constructivist model of radicalization and deradicalization: A conceptual and methodological proposal. *Frontiers in Psychology*, *10*, 412.

Gaines, B. J., Mondak, J. J. (2009). Typing together? Clustering of ideological types in online social networks. *Journal of Information Technology & Politics, 6,* 216–231. doi:10.1080/19331680903031531

Grauwin, S., & Jensen, P. (2012). Opinion groups formation and dynamics: Structures that last from non-lasting entities. *Physical Review Journal, 85*, 066113. doi:10.1103/PhysRevE.85.066113

Guilbeault, D., Becker, J., & Centola, D. (2018). Social learning and partisan bias in the interpretation of climate trends. *Proceedings of the National Academy of Sciences, 115*, 9714-9719. doi: 10.1073/pnas.1722664115

Guo, L., & Harlow, S. (2014). User-generated racism: An analysis of stereotypes of African Americans, Latinos, and Asians in YouTube videos. *Howard Journal of Communications, 25*, 281-302. doi:10.1080/10646175.2014.925413.

Harel, T. O., Maoz, I., & Halperin, E. (2020). A conflict within a conflict: intragroup ideological polarization and intergroup intractable conflict. *Current Opinion in Behavioral Sciences*, *34*, 52-57. doi:10.1016/j.cobeha.2019.11.013.

Harlow, S. (2015). Story chatterers stirring up hate: Racist discourse in reader comments on U.S. newspaper websites. *Howard Journal of Communications, 26*, 21-42. doi:10.1080/10646175.2014.984795.

Hoffman, A. J. (2011). The growing ideological divide over climate change. *Nature Climate Change, 1*, 195 – 196. doi:10.1038/nclimate1144

Holtz, R. (2003). Intragroup or intergroup attitude projection can increase opinion certainty: Is there classism at college? *Journal of Applied Social Psychology, 33,* 1922–1944.

Holtz, R., & Miller, N. (2001). Intergroup competition, attitudinal projection, and opinion certainty: Capitalizing on conflict. *Group Processes & Intergroup Relations, 4,* 61–73.

Holtz, R., & Nihiser, T. H. (2008). Relative deprivation, attitude contrast projection, and opinion certainty. *Group processes & intergroup relations*, *11*(1), 89-114.

Karlsen, R., Steen-Johnsen, K., Wollebæk, D., & Enjolras, B. (2017). Echo chamber and trench warfare dynamics in online debates. *European Journal of Communication, 32*, 257-273. doi:1 0.1177/0267323117695734

Himelboim, I., McCreery, S., & Smith, M. (2013). Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on Twitter. *Journal of Computer-Mediated Communication, 18,* 154-174. doi: 10.1111/jcc4.12001

Khan, M.L. (2017). Social media engagement: What motivates user participation and Consumption on YouTube? *Computers in Human Behavior, 66,* 236-247. doi:10.1016/j.chb.2016.09.024

JASP Team (2020). JASP (Version 0.12)[Computer software].

Kaakinen, M., Sirola, A., Savolainen, I., & Oksanen, A. (2020). Shared identity and shared information in social media: development and validation of the identity bubble reinforcement scale. *Media Psychology, 23*, 25-51. doi:10.1080/15213269.2018.1544910

Kelly, J., Fisher, D., & Smith, M. (2005). Debate, division, and diversity: Political discourse networks in USENET newsgroups. In *Online Deliberation Conference* (pp. 1-35). Stanford University.

Koleva, S. P., Graham, J., Iyer, R., Ditto, P. H., & Haidt, J. (2012). Tracing the threads: How five moral concerns (especially purity) help explain culture war attitudes. *Journal of Research in Personality, 46,* 184 – 194. doi:10.1016/j.jrp.2012.01.006

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4,* 863. doi: 10.3389/fpsyg.2013.00863

Livingstone, A., & Haslam, S. A. (2008). The importance of social identity content in a setting of chronic social conflict: Understanding intergroup relations in Northern Ireland. *British Journal of Social Psychology, 47,* 1-21. doi:10.1348/014466607x200419

Liu, C. C., & Srivastava, S. B. (2015). Pulling Closer and moving apart: Interaction, identity, and influence in the US senate, 1973 to 2009. *American Sociological Review, 80,* 192 – 217. doi:10.1177/0003122414564182

Iyengar, S., & Westwood, S. J. (2015). Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science, 59,* 690-707. doi:10.1111/ajps.12152

Maddison, S. (2012). Postcolonial guilt and national identity: Historical injustice and the Australian settler state. *Social Identities, 18,* 695-709. doi:10.1080/13504630.2012.709000

McGarty, C., Bliuc, A.-M., Thomas, E. F., & Bongiorno, R. (2009). Collective action as the material expression of opinion-based group membership. *Journal of Social Issues,* 65, 839 - 857. doi:10/1111/j.1540-4560.2009.01627.x

McGarty, C., Turner, J. C., Oakes, P. J., & Haslam, S. A. (1993). The creation of uncertainty in the influence process: The roles of stimulus information and disagreement with similar others. *European Journal of Social Psychology, 23*(1), 17-38.

Moran, A. (2011). Multiculturalism as nation-building in Australia: Inclusive national identity and the embrace of diversity. *Ethnic and Racial Studies, 34*, 2153-2172. doi:10.1080/01419870.2011.573081

Moscovici, S., & Zavalloni, M. (1969). The group as a polarizer of attitudes. *Journal of Personality and Social Psychology, 12*, 125-135. doi:10.1037/h0027568

Myers, D. G. & Bishop, G. D. (1970). Discussion effects on racial attitudes. *Science, 169*, 778-779. doi:10.1126/science.169.3947.778

Oksanen, A., Oksa, R., Savela, N., Kaakinen, M., & Ellonen, N. (2020). Cyberbullying victimization at work: Social media identity bubble approach. *Computers in Human Behavior, 109*, 106363. doi:0.1016/j.chb.2020.106363

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: measures of effect size for some common research designs. *Psychological Methods*, *8*(4), 434 - 447. doi:0.1016/j.chb.2020.10636310.1037/1082-989X.8.4.434

Pakulski, J., & Tranter, B. (2000). Civic national and denizen identity in Australia. *Journal of Sociology, 36*, 205-222. doi:10.1177/144078330003600205

Pariser E (2011) The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think (Penguin, New York)

Pauwels, L., & Schils, N. (2016). Differential online exposure to extremist content and political violence: Testing the relative strength of social learning and competing perspectives. *Terrorism and Political Violence, 28*, 1 – 29. doi:10.1080/09546553.2013.876414

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). *The development and psychometric properties of LIWC2007* Austin, TX: [LIWC manual].

Pennebaker, J., & Francis, M. E. (1996). Cognitive, emotional and language processes in disclosure. *Cognition and Emotion, 10*, 601- 626. doi:10.1080/02699939680079

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology, 77*(6), 1296-1312. doi: 10.1037//0022-3514.77.6.1296

Pennebaker, J.W., Booth, R.J., Boyd, R.L., & Francis, M.E. (2015). *Linguistic Inquiry and Word Count: LIWC2015.* Austin, TX: Pennebaker Conglomerates [www.LIWC.net].

Quattrociocchi, W., Scala, A., & Sunstein, C. R. (2016). Echo chambers on Facebook. *SSRN Electronic Journal*, doi:10.2139/ssrn.2795110

Reicher, S. D., Cassidy, C., Wolpert, I., Hopkins, N., & Levine, M. (2006). Saving Bulgaria's Jews: An analysis of social identity and the mobilization of social solidarity. *European Journal of Social Psychology, 36,* 49 - 72. doi:10.1002/ejsp.291.

Smith, L. G. E., Gavin, J., & Sharp, E. (2015). Social identity formation during the emergence of the Occupy movement. *European Journal of Social Psychology, 45,* 818-832. doi:10.1002/ejsp.2150

Smith, L. G., Thomas, E. F., & McGarty, C. (2015). "We must be the change we want to see in the world": Integrating norms and identities through social interaction. *Political Psychology, 36,* 543-557. doi:10.1111/pops.12180

Smith, L. G., McGarty, C., & Thomas, E. F. (2018). After Aylan Kurdi: How tweeting about death, threat, and harm predict increased expressions of solidarity with refugees over time. *Psychological Science, 29,* 623-634. doi: 10.1177/0956797617741107

Smith, L. G., Wakeford, L., Cribbin, T. F., Barnett, J., & Hou, W. K. (2020). Detecting psychological change through mobilizing interactions and changes in extremist linguistic style. *Computers in Human Behavior, 108,* doi:10.1016/j.chb.2020.106298

Stephan, W. G., & Stephan, C. W. (1985). Intergroup anxiety. *Journal of Social Issues, 41,* 157 – 175. doi:101111/j.1540-4560.1985.tb01134.x

Stephan, W. G., & Renfro, C. L. (2002). The role of threats in intergroup relations. In D. Mackie & E. R. Smith (Eds.), *From prejudice to intergroup emotions* (pp. 191–208). New York: Psychology Press.

Stephan, W.G., Ybarra, O., & Morrison, K.R. (2017). Intergroup threat theory. In T.D. Nelson (Ed.), *Handbook of Prejudice, Stereotyping, and Discrimination* (pp.43-55). New York: Psychology Press.

Sunstein, C. (2008). The Law of Group Polarization. In J. S. Fishkin & P. Laslett (Eds.), *Debating Deliberative Democracy*: Blackwell Publishing Ltd.

Taber, C. S., Cann, D., & Kucsova, S. (2009). The motivated processing of political arguments. *Political Behavior*, *31*(2), 137-155. doi:10.1007/s11109-008-9075-8

Tajfel, H., & Turner, J. C. (1979). An integrative theory of social conflict. *The social psychology of intergroup relations*, 2, 33-47.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC computerised text analysis methods. *Journal of Language and Social Psychology, 29,* 24 – 54. doi:10.1177/0261927X09351676

The Juice Media. (2017, January 25). *Australia Day Piracy Parody: You wouldn't celebrate September 11* [video file]. Retrieved from https://www.youtube.com/watch?v=UytdM-x3cv4

Thomas, E. F., McGarty, C., & Mavor, K. I. (2009). Aligning identities, emotions, and beliefs to create commitment to sustainable social and political action. *Personality and Social Psychology Review, 13*, 194 - 218. doi:10.1177/1088868309341563

Turner, J. C., & Tajfel, H. (1986). The social identity theory of intergroup behavior. *Psychology of Intergroup Relations*, *5*, 7-24.

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). Rediscovering the social group: A self-categorization theory. Oxford, England: Blackwell.

Vaccari, C., Valeriani, A., Barberá, P., Jost, J. T., Nagler, J., & Tucker, J. A. (2016). Of echo chambers and contrarian clubs: Exposure to political disagreement among German and Italian users of Twitter. *Social Media+ Society, 2*, doi:10.1177/2056305116664221

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). Rediscovering the social group: A self-categorization theory. Oxford, England: Blackwell.

Wojcieszak, M. (2010). 'Don't talk to me': effects of ideologically homogeneous online groups and politically dissimilar offline ties on extremism. *new media & society*, *12*, 637-655. doi:10.1177/1461444809342775

Yardi, S., & Boyd, D. (2010). Dynamic debates: An analysis of group polarization over time on twitter. *Bulletin of Science, Technology & Society*, *30*, 316-327. doi:10.1177/0270467610380011